The Ontario Test of English Achievement

The Final Report On Its Development

Council of Ontario Universities
Conseil des Universités de l'Ontario

1983

THE ONTARIO TEST OF ENGLISH ACHIEVEMENT

THE FINAL REPORT

ON ITS DEVELOPMENT

Principal Investigators: P. Evans
M. Holmes

Research: D. Jantzi
P. Haynes

The Council of Ontario Universities
Le Conseil des Universités de l'Ontario
130 St. George Street
Suite 8039
Toronto, Ontario M5S 2T4

July, 1983

83-7

This research was funded by the Ministry of Colleges and Universities, Ontario. The report reflects the views of the authors and not necessarily those of the Ministry.

TABLE OF CONTENTS

Section		Page
I	Background and Objectives	1
II	Process of Test Development	3
III	Subtest Characteristics and Properties	5
IV	Validity	10
V	Recommendations Concerning the Use of the Tests	22
VI	Other Recommendations	25

Appendices

A	Posttesting Written English
B	Report on Scorer Reliability
C	Descriptive Statistics for the Subtests

SECTION I

BACKGROUND AND OBJECTIVES

The English Evaluation Project was conceived in response to a request from the CONDAT sub-committee of the Council of Ontario Universities for the development of an evaluation procedure to be administered to students upon entering university. The variation in standards among high schools creates problems for universities, both when selecting students and when placing students into suitable programs and courses. It was thought that an evaluation procedure which would determine the language competence of their incoming students would be valuable in assisting universities in tailoring first-year English courses to meet the specific needs of their students.

The evaluation procedure developed for CONDAT was intended to be primarily a diagnostic instrument identifying areas of comparative strength and weakness in English. Decisions about how to use the results of the tests were made by the appropriate personnel within the participating universities to meet their own needs. Most universities planned to use the results to improve the match between student needs, as indicated by group performance, and the content of first-year English courses. Individual scores could be used to direct students to remedial English courses if performance was not at a level judged acceptable by the individual university. Where universities varied the content of first-year courses to deal with specific areas of language use, students could be streamed into the courses, or sections within courses, which would be most beneficial for them. Another potential use of individual results would be to identify those students who would benefit most from an English course in which more students wished to enrol than could be accommodated by the resources available. Finally, the results could also be used to determine whether a significant number of incoming students lacked sufficient competence in their use of language to make possible their successful completion of the university program. If that were the case, new courses addressing problem areas could be developed for students demonstrating need. All participating universities indicated their desire to act on at least one of the options outlined above.

Since the use of English includes more aspects than could be tested reliably in the time available, the principal investigators, in consultation with the initiating universities, identified four components believed critical to success in university. The components tested are reading ability, writing

ability, ability to plan and develop a theme, and competence in the use of mechanics of language. The subtests developed for each component are discussed in detail later in this report.

The objectives of the test development were:

(1) to develop valid, reliable measures of the identified areas of competence;
(2) to develop a test that can be administered economically;
(3) to enable university teachers to make useful recommendations with respect to the initial placement of students in different English courses;
(4) to enable university teachers to diagnose specific areas where competence is lacking;
(5) to enable students to obtain remediation in specific areas of weakness;
(6) to provide tests that can be used in contexts other than those in which they are initially developed: e.g. other universities with similar problems;
(7) to provide measures of the effectiveness of remedial progams.

SECTION II

PROCESS OF TEST DEVELOPMENT

In the fall of 1980 Wilfrid Laurier and Laurentian Universities administered a language test (developed primarily by WLU) and a test of writing composition to some of their first-year students, many of whom were volunteers. Peter Evans had provided advice and assistance in item development and in the analysis of test results. When the CONDAT project was established, Carleton University and, later, the University of Guelph joined Wilfrid Laurier and Laurentian as participants in the project.

The test development began in the winter of 1980/81 with teams of university and secondary-school teachers in co-operation with the principal investigators and the research officer associated with the project, Doris Jantzi. The following were team members: Michael Wendt, Robert Lovejoy, Larry McDonald and Laura Groening of Carleton University; Judy Jewinski, Meg Hancock and John Chamberlain of Wilfrid Laurier University; Pamela Steel and Margaret Hundleby of the University of Guelph; Hoyt Greeson of Laurentian University; Michael O'Hagan of the Sudbury Board of Education; Yvonne Benton of the Frontenac Lennox & Addington RCSS Board of Education; Harry Corrin of the Carleton Board of Education; Pat Haynes and Marvle Bell of the OISE Midnorthern Centre.

Preliminary versions of the subtests were tried out and calibrated in a number of settings, including the Boards of Education for Carleton, Lincoln County, Sudbury, Waterloo County, Wellington County RCSS and York County, as well as the participating universities. The results of the screening trials were used to revise, delete or replace items for the preliminary testing done by the universities in September of 1981. Two parallel forms of the test were developed so that one form could be used as a pretest and the other as a posttest at the end of the first-year English course dealing with the problem areas identified.

The preliminary test was used differently in the various settings. Carleton University administered the test to 80 student volunteers in their commerce program. At the University of Guelph 1700 students from all programs were required to write the test. Similarly, Wilfrid Laurier University required 1200 students to write the test. Laurentian University administered the test to 350 student volunteers in nursing, physical education and engineering programs. At WLU test scores were used to determine enrolment in the first-year writing course by admitting the bottom 700 on a rank-ordered list. At Carleton test

results were used to assign students to different English courses. At Laurentian test scores were used primarily to assist in advising students concerning their selection of courses and, in some cases, to exempt students from an ESL course if their scores were higher than a predetermined point. Guelph was interested primarily in using the results to determine if a remedial English program was needed in their university program.

In December Guelph and Wilfrid Laurier administered the alternate form of the test as a posttest to students who had completed a first-year writing course in the fall term. Carleton and Laurentian administered the posttests in March at the end of their writing courses. The results of an item analysis of the preliminary testing in 1981/82 were used to revise the tests for the final version submitted with this report.

The participation of the universities varied, with some involved to a greater extent than others. Two of the universities set up training sessions for their essay scorers and subsequently had greater reliability among their scorers. At several universities professorial assessments were given for the posttest group as a validity check on the tests.

Throughout the project an Advisory Committee, consisting of Hoyt Greeson of Laurential University, Ian Cameron and Robert Lovejoy of Carleton University, Michael Moore of Wilfrid Laurier University, Gerald Rubio and James Harrison of the University of Guelph, Naz Deen of the Carleton Board, John Fitchett of the Ottawa Board, Michael O'Hagan of the Sudbury Board, Frank Pendergast of the Waterloo County Board, May Simpson and Sheila Paterson of the Wellington County RCSS Board and Ian Hamilton, representing COU, provided general direction and advice to both the principal investigators and the participating universities.

SECTION III

SUBTEST CHARACTERISTICS AND PROPERTIES

1. Reading Comprehension

The Reading Comprehension subtests (for Forms A & B) consist of a series of multiple-choice questions with four options, one of which is correct. A period of 45 minutes is allocated for the subtests. There are three passages in each subtest, 27 questions on Form A and 25 on Form B. Of the six passages, four are based on economics, one on an excerpt from a work of fiction and one on urban geography. A knowledge of economics may be of a little assistance, but all the questions can be answered directly from the passages. None of the test developers was an economist.

The passages are quite dense. They are at a level frequently met by students in their years at university. There are no significant literary or technical hurdles. Although mean scores in the sample populations ranged from 45% to 52%, it would be unfair to reach a conclusion that one-half or even one-third of the university population is unable to read adequately. The questions are, quite deliberately, difficult. Some questions require shades of judgment over which there will be disagreement among reasonable and educated people. Such conceptually demanding questions are unlikely to achieve perfect discrimination.

The question of validity is addressed later in this report. We do not address at this point, therefore, the question of whether the subtest measures what it is supposed to be measuring. The question of reliability is also important. We are concerned about reliability because we wish to know whether the scores are reasonably precise. The coefficient of reliability tells us whether we are in fact measuring something that is reasonably stable; validity tells us whether we are measuring what we want to measure. The reliability of the subtests has varied according to the site of administration and the form of the subtest. The coefficients of reliability are in the 0.60 to 0.65 range. Some minor adjustments have been made to the tests since they were last used but we do not anticipate they will significantly improve the reliability coefficients.

The following considerations should be taken into account in assessing the usefulness and the reliability of the Reading subtest:

(a) it is very difficult to discriminate precisely among university students as they form a selected sample of the population;

(b) the subtest is only a part of a larger test and major decisions should not be based on a subtest with a reliability of 0.60;

(c) the reliability coefficients have been calculated within, as distinct from across, university populations as it is assumed that the tests themselves will be used within rather than among universities; reliability coefficients might be a little higher across universities as the population of Ontario universities is probably a little more diverse than the population samples within the universities within the project.

2. Thesis Development

The Thesis Development subtest consists of a series of multiple-choice questions with four options, one of which is correct. A period of 25 minutes is allocated for the subtest. In each subtest, questions (21 in Form A, 22 in Form B) are based on three simulated theses. Questions demand the identification of appropriate first and last sentences, essay outlines, supporting arguments, theses, and topic sentences. Occasionally, some knowledge of the simulated essay topic may be helpful but the topics were chosen to eliminate the advantage of specialist knowledge (except in the area of essay structure) as much as possible.

The opinion of the principal investigators, not entirely shared by all members of the Advisory Committee, is that the ideas of thesis development could be effectively taught and thus scores on the test improved over a period of time. The universities themselves did indeed provide special language instruction to some students who took the tests in September of 1981, but it is unclear to what extent the instruction focussed on thesis development as distinct from more general skills of language and writing.

As thesis development is not considered a focus for teaching and learning in school (to the extent that reading comprehension and writing are), it is more than usually difficult to comment on the achievement level. Mean scores on the subtest are in the 60-65% range and the reliability coefficients are in the 0.60-0.65 range.

Some critics will undoubtedly believe that the subtest asks inappropriate and misguided questions. They will believe that writing flows naturally and

should not suffer obstruction as a result of rules of format, structure, etc. This issue was debated at length by the task developers, including teachers and university professors of English, and by the Advisory Committee. There was consensus, virtual unanimity, that some basic skills in writing development should be learned even though many advanced writers may go on to ignore them with impunity. The argument is not that no one can learn to write effectively without learning the structural rules; it is not that all good writing follows a set of implicit or explicit rules; rather is it that the majority of university students benefits from knowing how to structure a formal essay.

3. Language

Each of the two forms of the language test consists of 40 multiple-choice items to be administered in a period of 25 minutes, an amount of time that appears sufficient for most students to complete the test. Each test comprises four subtests: Spelling, with some of the chronic spelling problems as targets, such as ei/ie, consonant preceding a suffix, the apostrophe; Grammar, chiefly problems with the verb and conjunctions; Diction/Usage, with half the items testing word confusion and half appropriate style (register, cliché, redundancy); and Sentence Structure. The two forms were carefully matched, item for item, for the target under test and balanced for anticipated difficulty on the evidence from screening trials.

Both forms performed well with reliability estimates in their various administrations ranging from 0.63-0.72. The two forms were compared for difficulty in a special administration to 111 Grade 12 and 13 students in the Carleton Board (alternative forms administered in random order) and Form B was found more difficult than Form A by almost exactly two raw score points.

Subsequently, following item analysis, review by the Advisory Committee and the consequent revision of a few items, several items were exchanged in an effort to make the two forms equivalent in difficulty.

In the Carleton Board administration the statistics for the original Forms A and B were as follows:

	Mean	Standard Deviation	Estimated Reliability	SEM
Form A	20.6	4.8	.64	2.86
Form B	18.7	4.9	.65	2.82

Administration in all university settings produced mean scores in the 45-50% range. It must be remembered that the language test, unlike the others, concentrated entirely on problems, and so very high scores are not expected. Some members of the Advisory Committee have demurred in this judgment of the tests as fairly representing an estimate of language performance. In response it will be seen in Section IV that statistically the language test behaved best of the four as a stable measure.

The lertap program used to score the subtests in objective format calculated a reliability coefficient for the total test as well as for the individual subtests. As noted above, the subtest reliabilities ranged from 0.63 to 0.72, varying with the subtest and setting. The relatively low reliability coefficients are attributable to both the lack of variation in the population tested and the small number of items on two of the subtests. The reliability coefficient for the total test (Reading, Thesis Development and Language) was higher, ranging from 0.80 to 0.83 in the various settings. The higher total test reliability reflected the positive correlations of the subtests with each other and with the total test.

4. Writing

The topic for the Written English subtest is expository (point of view or argument), the mode recommended in the Ontario Ministry of Education Senior English Guideline (p. 94) as the one perceived as most important for post-secondary education. Students were allowed 50 minutes to sketch an essay in rough and to recopy it onto treated paper producing multiple copies for scoring.

From a short list of about ten stimuli created or selected by the writing team, two were finally selected as Forms A and B. Late in the summer of 1981 it became evident that a third stimulus would be required at Guelph, as there were to be three testing occasions in September and one could not expect to maintain confidentiality of topic. This third stimulus was not pretested but did work quite satisfactorily. Its weakness is its topicality. A report concerning the RCMP had just been released and so the topic had a relevance in September 1981 that it would not have had a year later. We are therefore recommending only Form A and Form B.

Essays were scored holistically on a scale of 1-10 by a minimum of two scorers. Ten essays were buried for scoring by all, and scorers calibrated to bring their scores to the same mean. Calibrated scores on the 1-10 scale were

averaged and scaled 100-200 with a mean of 150 and a standard deviation of approximately 20.

In two universities (Guelph and Laurentian) Co-Investigator Evans trained the scorers for the September administration. The benefits of initial training were evident in the analysis of scorer behaviour. The training package used is provided as an appendix with the recommendation that it be reproduced for use in future training of scorers.

Emphatically we recommend a minimum of two (preferably three) scorers, with buried set and calibration, and a training session for scorers.[1]

---

[1]Appendix A discusses the special problem of pre and posttesting students in the area of written English. In Appendix B scorer reliabilities are given for three testing occasions. The reliabilities which are normal for holistic scoring illustrate the great importance of using appropriate methodology and scorer training.

SECTION IV

VALIDITY

One of the most important factors in the consideration of a test is its validity. In Section III we dealt with the question of reliability of subtests: i.e. how accurately do subtests measure whatever it is they are measuring?

Validity is of central importance and is also a significant problem. Even if we agree that writing is an important skill (and not everyone does), how do we know that our Written English and Language subtests in fact measure writing? We may compare two tests, but by what criterion is one a judge of the other?

We chose the following methods of checking validity: (1) involve school and university teachers in the test development so that tests will measure what practising teachers think the various subtest domains really mean; (2) have an advisory committee examine the subtests and see if they appear to be measuring what they are intended to measure; (3) compare the relative achievement of students on the different subtests, on the assumption that there should be some relationships among the subtests but a lesser one than between two forms of the same subtest: i.e. the subtests should be measuring some of the same things, but they should also be measuring some different things; (4) compare subtest scores with the estimates of the same students' university teachers; (5) compare improvement on subtests during instruction with the university teachers' estimates of improvement; (6) compare subtest and test scores with other measures of achievement: e.g. Grade 13 marks and university grades.

The first two approaches have already been described in Section II of this report which deals with test development. A reasonable cross-section of school teachers and university teachers believes the tests to be valid. In many ways, such a measure of validity is the most important indicator, particularly if the view of the teachers concerned is shared generally by school and university teachers of English throughout the province.

The problem with all the remaining indicators is that they themselves are imperfect; if they were good measures we would not need a test. Thus the following validity assessments must be treated with great care. If the relationships prove to be very strong, then a critic may argue that we do not need the tests because they add little to our current knowledge. If the relationships are weak, the critic may argue that the tests are of little value as they are

not clearly related to the other tried and tested criteria. By the same token, if the relationships are weak, the principal investigators can claim that the tests inevitably measure different things from the other measures. Nonetheless, we do believe that an examination of the relationships can give us a very general picture of the tests' validity. The following discussion is based on data from the two universities for which we have relatively complete data -- Wilfrid Laurier and Guelph. Wilfrid Laurier University experienced some problems in the September administration of the test, particularly with respect to the scoring of Written English, so we would generally advise that greater weight be given to the Guelph data.

Question

How well do the pretests correlate with each other?

Table 1

Correlation Matrix of Pretests in Writing, Reading, Thesis Development, and Language - Wilfrid Laurier and Guelph

	Pre-R		Pre-T		Pre-L	
	WL	G	WL	G	WL	G
N	227	171	227	171	227	171
Pre-W	-.07	.37**	.12	.40**	.14	.59**
Pre-R			.30**	.50**	.25**	.42**
Pre-T					.23**	.40**

** Indicates $p < .001$

It must be borne in mind that the reliability coefficients of the subtests are in the 0.60 range. Therefore, very large intercorrelations are impossible. We would hope that the Language and Written English subtests would be strongly related: they purport to measure many of the same things and, in the case of Guelph, which university has provided the cleanest data, this in fact proves to be the case. The number of students on which the coefficients are based varies from approximately 170, in both universities, to a very much smaller number when we deal with professorial estimates, which were difficult to obtain. Hence, the magnitude of the coefficient required for statistical significance varies considerably. With a large n, a comparative small coefficient will be statistically significant. With a very small n, a coefficient as large as 0.25 may not be

statistically significant. Table 1 shows reasonable levels of relationships, with the exception of Writing in Wilfrid Laurier.

Question

Does performance on the pretests predict performance on the posttests?

Table 2

Correlation Table Showing Relationships Between Pre and Posttests

	Post-W		Post-R		Post-T		Post-L	
	WL	G	WL	G	WL	G	WL	G
N	227	168	211	171	213	171	222	171
Pre-W	.17*	.52**	.04	.41**	.10	.40**	.17	.47**
Pre-R	.13	.32**	.44**	.44**	.36**	.38**	.29**	.35**
Pre-T	.16	.37**	.36**	.45**	.37**	.46**	.25**	.40**
Pre-L	.24**	.49**	.34**	.38**	.17	.33**	.55**	.51**

* Indicates $p < .01$
** Indicates $p < .001$

If each subtest is reliable, there is likely to be a strong relationship between performance on Form A and performance on Form B. Furthermore, if the subtests really measure different things -- as distinct from generalized verbal intelligence, for example -- then we would expect the relationships between the pre and posttest forms of the same subtest to be stronger than the relationships between pre and posttest forms of different subtests. Once again, if we depend mainly on the more credible Guelph data, we note that in the case of Writing the pretest correlates virtually as well with Language as with Writing and the relationships with Reading and Thesis Development are almost as high. In the case of Reading the relationship with Thesis Development is almost as high as with Reading itself, and the same is true, *vice versa*, in the case of Thesis Development. As for Language, it is related almost as strongly to Writing as to the second form of Language.

If we had highly reliable subtests, and if we tested and retested on consecutive days, we would hope to obtain correlation coefficients in the 0.8 to 0.9 range. But the subtests are not highly reliable and the posttest took place several months later, and following a period of instruction designed to improve performance. Thus the results can hardly be described as surprising. It is

not clear, however, that the Language and Writing subtests really measure very different abilities; and the same may be said of the Reading and Thesis Development subtests. Furthermore, either students with strength in one subtest domain normally have the same strength in the other subtest domains or the four subtests are measuring quite similar abilities.

Question

Do professorial pre-assessments correlate well with pretest scores?

Table 3

Relationships Between Professorial Assessments and Pretest Scores

	Pre-PA W		Pre-PA T		Pre-PA L		Pre-PA R	
	WL	G	WL	G	WL	G	WL	G
N	196	53	196	53	196	53	32	30
Pre-W	.15	.38*	.17	.34	.12	.53**	.20	.37
Pre-R	.15	.49**	.15	.21	.18	.45**	-.04	.39
Pre-T	.21*	.39*	.15	.29	.18	.32	.28	.23
Pre-L	.27**	.25	.29**	.30	.26**	.36*	.14	.05

* Indicates $p < .01$
** Indicates $p < .001$

The number of students on which the statistics are based is quite small in many cases. The size of the coefficient required for statistical significance is thereby increased. The relationships, though generally positive, are insufficiently strong to permit the drawing of firm conclusions. This is not surprising, as professors objected to the exercise on the grounds that they could not possibly assess students in these four areas so early in the term. We feel that this very fact, confirmed by the results of the assessment, underlines the usefulness of the test. At the end of September, university teachers believe themselves, correctly, to be incapable of making such assessments. If the assessments are to be made, some source outside the university is necessary.

Question

Did the students improve on the subtests between September and December? (See Table 4)

In one university Form A was used in September and Form B in December. In the other university the use of the forms was reversed. As we cannot be sure that the level of difficulty of the two forms is identical, it seemed most

TABLE 4

Changes in Scores on the Subtests between September, 1981 and December, 1981

	Pre-T		Post-T		Pre-L		Post-L		Pre-R		Post R	
	WL	G	WL	G	WL	G	WL	G	WL	G	WL	G
N	227	171	213	171	227	171	222	171	227	171	211	171
Mean Scores (%)	61.4	55.2	66.2	58.6	42.8	42.8	51.8	43.5	47.6	47.8	51.9	48.8
S.D.	2.6	3.1	2.6	3.1	4.2	5.5	4.6	5.5	3.0	3.6	3.5	3.3
Improvement in Test	-	-	4.8	3.4	-	-	9.0	0.7	-	-	4.3	1.0
Improvement in Metric of Pretest S.D.	-	-	1.9	1.1	-	-	2.1	0.1	-	-	1.4	0.3
Average Improvement in S.D. Units	-	-	1.5		-	-	1.1		-	-	0.9	

sensible to assess each university's 'improvement' in performance in terms of a percentage score expressed in the metric of the pretest standard deviation and then show the average of the two universities. Improvement in Writing is difficult to assess as the stimuli and scorers were different. These estimates are only approximate.

It is noticeable that improvement is shown in the three subtests. As it is generally quite difficult to show improvement from instruction among mature students in the area of English, we must be cautious in our intrepretation. One interpretation of the improvement is that the students taking the posttests are poor students whose pretest scores are merely showing a regression towards the mean. There is indeed some support for this supposition. In Thesis Development, for example, the mean score of the Guelph students writing the test in September was 66.57%, considerably higher than the 52.73% recorded in Table 4 for the sample who took the posttest in December. In Reading, however, the mean scores were almost identical (but the improvement was also less in this case). In the case of Language, the mean score of the entire September population in Guelph was 48.93%, once again higher than the sample in Table 4. A second explanation for the improved score is that the students were more familiar with the tests by the time of the second administration. While that cause cannot be ruled out, it appears unlikely that university students would be so unused to tests and so much affected by their unfamiliarity. A third explanation is that the students were somewhat disoriented by the orientation process in early September. It is by no means impossible that some students took the tests the morning after. A fourth possibility is that the students' English improved during the period as a result of the specific instruction provided them. A very slender piece of supportive evidence for this explanation is the fact that improvement was greater in Language and Thesis Development than in Reading. The university programs are directed towards writing skills rather than reading skills. Thesis Development, we believe, is the area most sensitive to instruction. However, we have no knowledge that instruction did in fact concentrate on Thesis Development problems more than on more general problems found in the Language subtest.

Question

What is the difference between professorial assessments at the beginning and end of term? (See Table 5)

TABLE 5

University Professors' Assessments During the Pretest and Posttest Periods

	Pre-PA W		Post-PA W		Pre-PA T		Post-PA T		Pre-PA L		Post-PA L		Pre-PA R		Post-PA R	
	WL	G	WL	G	WL	G	WL	G	WL	G	WL	G	WL	G	WL	G
N	196	53	196	53	196	53	196	53	196	53	196	53	32	30	46	53
Mean Score	2.8	2.7	3.1	3.1	2.8	2.4	3.1	3.0	2.8	2.8	3.0	2.9	2.9	3.3	2.9	3.0
S.D.	.9	.8	.8	.8	.8	.9	.9	.8	.9	.8	.8	.8	.8	1.0	1.0	.7
Improvement in Test	-	-	.3	.4	-	-	.3	.6	-	-	.2	.1	-	-	.0	-.3
Improvement in Metric of S.D.	-	-	.33	.50	-	-	.38	.67	-	-	.22	.13	-	-	.0	-.3
Average Improvement in S.D. Units			.42				.53				.18				-.15	

It is not surprising that professors would generally evaluate their students as improving, but it is noteworthy that they do not see any apparent improvement in reading. Presumably the course provided by the university is directed more at skills in writing than at skills in reading. There is a discrepancy between the perceived improvement in thesis development by professors and the actual improvement, according to the tests, in language. Does this mean that the professors believe they are teaching thesis development more than written English?

It is difficult, then, to come to any firm conclusions on the subtests from Tables 4 and 5. Both measures do show improvement but there are other explanations for the improved scores than actual, enhanced competence. The professors (i.e. the very small number responding to this part of the study) see improvement in Writing, Thesis Development and, slightly, Language. The tests suggest improvement in Thesis Development and Language. There is nothing in the data to refute a hypothesis suggesting that the subtests are valid and that some improvement in some domains was achieved over the term. On the other hand, there are other explanations for the findings.

Question

What is the relationship between professorial assessments of improvement and tests' measures of improvement?

Table 6

Correlation Table Showing Relationships Between Professorial and Subtest Assessment of Improvement in Performance

	DIF W		DIF T		DIF L		DIF R	
	WL	G	WL	G	WL	G	WL	G
PRG.PA. W	.00	.26	.04	.05	.22*	.05	.16	.13
N	195	53	195	53	195	53	195	30
PRG.PA. T	.01	.31	.04	-.03	.19*	.05	.07	.16
N	196	53	196	53	196	53	196	30
PRG.PA. L	.01	.35	.07	.00	.16	.20	.08	.11
N	196	53	196	53	196	53	196	30
PRG.PA. R	-.06	.26	-.10	.05	.20	.07	.25	.04
N	46	53	46	53	46	53	46	30

* Indicates $p < .01$

The data are insufficiently clear to support firm generalizations. That is to say, the difference between pre and posttest scores is not sufficiently related to professorial assessments of improvement during the course for us to

state that the tests and university teachers see the same improvements in the same people. However, as we have already noted, university teachers are understandably unable to assess students at the beginning of the term so it should not be surprising if they have difficulty in assessing improvement, even if we assume there was in fact some improvement to assess. The small number of cases on which we have data also obscures any possible relationship. It is possible that there might be some significant relationships if we had more data (e.g. between writing scores and professorial assessments of progress in language).

Question

What is the relationship between student performance on the test and subtests and other indicators of student performance? (See Table 7)

This approach to validity is second in importance only to the content validity checked by school and university teachers. If the test is a valid measure of competence in English, and if university courses in English and other subjects really require such competence, then there should be some relationship between test performance and performance expressed in university grades. If the Grade 13 marks predict university performance as well as or better than the test, then a case can be made that the test is superfluous -- if its purpose is merely to advise students with respect to program options. Of course, if the test is to be used as a universalistic yardstick: i.e., if "passing" the test is to be mandatory or if it is to be used for selection purposes, for university, for courses or programs, then Grade 13 marks are not self-evidently a useful substitute even if they do predict university grades better than the test does. There are two reasons for this assertion. Firstly, tests probably measure somewhat different sets of qualities from those measured by teachers; hence, different teachers' assessments are likely to be similar to each other and different tests' assessments are likely to be similar to each other. Thus, even if Grade 13 marks are better predictors of university marks than are marks from tests, one can't assume *ipso facto* that teachers' marks are the better measure. Secondly, if entry (to a university or to a program) is to be permitted and denied according to an assumed level of competence, there is much to be said for having all applicants submit to the same procedure, even if that procedure is far from being perfect. There may well be systematic bias in Grade 13 marks if some teachers inflate grades much more than do others.

TABLE 7

Relationships Among the OTEA Tests, Subtests and Other Indicators of Performance

	G 13 Eng.		G 13 Av.		Univ. Eng.[1]		Univ. Av.	
	WL	G	WL	G	WL	G	WL	G
PRE OTEA	.39**	.44**	.44**	.39**	.28**	.34**	.34**	.34**
N	179	66	187	58	184	157	179	157
POST OTEA	.39**	.06	.40**	.14	.32**	.44**	.37**	.35**
N	179	66	184	58	185	157	179	157
G 13 Eng.	--	--	.63**	.77**	.39**	.25	.42**	.44**
N			179	58	179	66	179	66
G 13 Av.	--	--	--	--	.39**	.25	.63**	.62**
N					184	58	179	58
PRE W	.26**	.36*	.28**	.26	.16	.38**	.16	.34**
N	179	66	184	58	184	152	179	152
PRE R	.14	.26	.26**	.10	.11	.33**	.24*	.31**
N	179	66	183	58	183	152	179	152
PRE T	.38**	.49**	.22*	.20	.18	.29**	.22*	.40**
N	179	66	183	58	183	152	179	152
PRE L	.38**	.49**	.43**	.34	.36	.38**	.40**	.37**
N	179	66	183	58	183	152	179	152
PST W	.17	.35*	.22*	.18	.30**	.52**	.17	.34**
N	179	66	183	58	183	149	179	149
POST R	.25**	.34*	.23*	.24	.24*	.29**	.27**	.33**
N	171	66	171	58	171	154	171	154
POST T	.19	.28	.23**	.19	.24**	.40**	.22*	.38**
N	173	66	173	58	173	154	173	154
POST L	.38**	.56**	.47**	.41*	.45**	.44**	.42**	.43**
N	180	66	80	58	180	154	179	154

* = $p < .01$
** = $p < .001$

[1]University English is the English Writing Course.

From a practical point of view, then, this question is of fundamental importance. If Grade 13 marks predict university marks much better than do the test scores, some people will undoubtedly argue that the tests are a waste of time and money -- and the principal investigators would sympathize with that position if the tests were only to be used for general guidance.

As Table 7 shows, the OTEA pretest is quite well related to Grade 13 English, university English and the university average. The posttest may be slightly, but most probably not significantly, more strongly related to the final university grades and less strongly to the school grades. The relationships generally reinforce the validity of the test. The Grade 13 English mark performs about as well as the test in predicting performance in university English and slightly better in the prediction of overall university performance.

Thus it would be difficult to justify the use of the OTEA test if one's only purpose was to advise students whether or not they should take a particular English course in university. However, a strong case can be made for the valid use of the OTEA test in situations where actual decisions about entry or course placement are to be made on a mandatory basis. Simply put, we can speculate that the Grade 13 mark is "better" than the OTEA score to the degree that teachers know students quite well and have frequent occasion to evaluate them. The Grade 13 mark is the sum of all those occasions. The OTEA test's advantage offsets but does not override that advantage. The OTEA test is "better" in that it provides an evaluation in the same format at the same time for all students, whereas those same students have received widely varying treatment from various teachers.

The subtests perform similarly -- none stands out as being indispensable and none appears dispensable. The Language subtest appears to perform the most consistently. In fact, its relationship with the other variables is as strong as that of the entire test as it proves the most consistently reliable.

Thus, if the test is to be used for decision-making purposes, the Language subtest should certainly not be deleted. If we compare the relationships between the subtests (after instruction) and university English and average grades, Written English stands out as the subtest that most clearly differentiates specific from general performance. At Guelph there is a correlation coefficient of 0.52 between the posttest Written English score and the university mark for the course. That coefficient is higher than the coefficients for the other subtests. Very tentatively, one might argue that the Written English subtest

reflects specific skills in English more and generalized intelligence less than do the other subtests. If the test is to be used as a measure of English competence, then, there is argument on statistical as well as pedagogical grounds for including the Written English component. If the test were to be used as a general criterion for university entrance, and if the pedagogical effects on high schools were to be ignored, it is possible that the OTEA test without the Written English component would perform about as well as the complete test. The remaining two subtests are useful but not absolutely statistically essential. Nevertheless, the more complete the test, the more its performance can be relied upon in varying circumstances with various student bodies. The principal argument for inclusion of all the subtests is one of validity rather than reliability. Thus, Reading and Thesis Development appear to be quite strongly related to general success in university. Presumably testing in English ought to be as much concerned with success in the university as a whole as it is with success in an English course.

In sum, we have taken two approaches to the problem of validity. The first approach is to ensure that the test evaluates those aspects of competence in English that school teachers believe to be important. There will never be unanimity on so controversial a topic but we believe a reasonable consensus has been reached that the tests represent what should be tested if one is going to test. The second approach is to examine statistically the relationships between our measures and the measures of others in schools and universities. The problem here is that there is no agreed standard: if there were, we would not be designing a test. However, in general, the OTEA test has performed as well or better than one might reasonably expect. It is a single-shot, universalistic test lasting three hours but it predicts university performance in English as well as assessments based on the whole year's experience in school.

SECTION V

RECOMMENDATIONS CONCERNING THE USE OF THE TESTS

In 1981 each university employing the tests made use of the results in slightly different ways: for advising students, for exemption of students from a basic English course, or for establishing (in a context where there were more students wishing to take the basic course than there were places) which students would be allowed to take the basic course. In all cases, the battery was a post-admission test; the score was not used to regulate admission to the university, to refuse admission to a particular program or to assist in determining the award of scholarships and bursaries. Also, in some settings the test was voluntary; in others it was required.

Under the mandate from CONDAT, the Co-investigators developed the battery as a post-admission test to assist universities in placing (or advising) students deficient in language competences considered important for university success. Alternate forms were developed for those universities wishing to post-test students who chose or were placed in basic English programs, as part of the measure of the student's improvement (and secondarily as a measure of the effectiveness of the program itself).

Extension of the use of the battery beyond the intended purposes in its development requires careful consideration.

1. We do recommend that all four subtests be employed, for the reasons outlined in Section IV of this report. Elimination of any component reduces validity and, to some degree, reliability. It is important to recall that each subtest is intended to measure somewhat different aspects of language competence deemed important for university success.

2. The proposed weighting of the subtests is a recommendation of the Co-investigators sustained by the Advisory Committee. The recommended weighting is the estimate of the relative importance of the subtests in judging the students' overall performance in language.

Proposed Weights for the Subtests

	Maximum Score	Weight
Reading	27*	2
Thesis Development	21*	1
Language	40	1
Written English	200**	1
Total	315	

*These scores are for Form A. The maximum scores for Form B are 25 and 22 respectively.

**The Written English score is on a 100-200 scale.

3. For pretesting writing, as noted in Section III, holistic scoring is recommended with a minimum of two scorers (preferably three), burial of a set of essays for scorer calibration, and a training session for scorers. The scale of 1-10 (an even-numbered scale) appears quite satisfactory. Essays on different topics should not be scored together.

In posttesting, where only the weaker writers are re-tested, holistic socring is probably infeasible. We have tentatively recommended (see Appendix A) a 3- or 4-point classification of the writing based on a closer assessment.

4. Conditions of testing: Current practice varies, but nowhere is it ideal. In those settings in our experience students had little or no advance information about the battery's contents or purposes. Nor did there seem to be a clear policy concerning who must take, ought to take, or might choose to take the battery, a problem exacerbated by the tight scheduling of the battery into a single administration over a three-hour period, usually for very large groups, in registration week. Some confusion, both concerning simply appearing for the test and in the test's administration, was inevitable.

A further pressure, as a consequence of registration-week testing, arose from the need in most settings to have scores almost immediately available: extreme pressure on computer services at a critical period and more extreme pressure to get the essay scoring done very rapidly.

As basic recommendations, the following suggestions should be conveyed by COU to anticipated users of the battery:

(1) policy within the university should first be established concerning who should/must write and why;

(2) if testing is to be conducted in registration week, applicants should receive prior information concerning the contents of the battery, the purposes of the testing, whether or not they must write or ought to write, and directions concerning dates, times and locations of administration;

(3) if the battery can be administered prior to registration week, this is a desirable practice; again, applicants should be fully informed as in (2);

(4) early planning and coordination are required: liaison should be developed with the Registrar's Office, with departments or divisions within the university, with computer services, and with planners of registration-orientation activities; within the English department, arrangements should be made in advance for selection of scorers and payment (if appropriate), training must be carefully planned;

(5) early planning and coordination may permit use of pre-coded student number labels for all subtests, a practice highly advantageous in compiling student data quickly and smoothly;

(6) with reference to (2) and (3), a brief Student Guide giving the purpose of the battery and sample items should be prepared.

SECTION VI

OTHER RECOMMENDATIONS

1. Analysis of the Writing (Sept. 1981)

The Advisory Group has requested that a sample of the writing holistically scored in September 1981 be analysed and a supplementary descriptive report be prepared.

Part of the Advisory Group's concern arose from discomfort with the language test and the relatively poor scores obtained, their concern being that performance data there give an incorrect impression of student competence as exhibited in actual writing. It would be fairer, in their view, if some evidence of the actual frequency of errors in different categories were provided. On the positive side, however, there are clear advantages in preparing a descriptive report on the writing. First, there is a good deal of interest on the part of university personnel, the teaching profession and the public at large concerning the characteristics of the writing of the university-bound student. Second, a reasonably complete descriptive report of the writing produced may be of value in the development or revision of first-year writing programs. Third, some matching of performance (error counts, analytic scoring) to obtained scores with exhibitions of sample compositions could be very helpful in establishing, on a more informed basis, appropriate cutting points or defining a reasonable minimum standard in writing for university entrance.

Such an analysis of, say, 100 essays from Form A and Form B, September, 1981 administration would include analytic scoring on, possibly, four criteria and error counts in a number of fairly simple categories by a team of eight scorers who have some previous experience in this type of analysis (OISE Writing Evaluation Project, Interface Study, etc.). Some aspects of primary trait scoring or measures of syntactic maturity might also be attempted.[1]

---

[1] P. Evans, "The Evaluation of Writing in Ontario: Grades 8, 12 and 13", Review and Evaluation Bulletin, Vol. 1, #2, Ontario Ministry of Education, 1979.

2. Development of Additional Forms of Subtests

If the number of users and uses were to increase, two current forms of each subtest would be insufficient to ensure reasonable confidentiality and flexibility. The problem of confidentiality is increased in those settings where two or more sittings are required to accommodate large numbers of students.

The preparation of two additional forms of each of the Reading, Thesis Development, and Language subtests is recommended. As well, all forms developed should be reasonably exactly calibrated for equivalence (with probable introduction of standard scores), hence interchangeability.

The provision of four forms for each of these subtests will increase very considerably the attractiveness of the OTEA battery for potential users.

The case of Writing is particularly urgent. Two stimuli are clearly insufficient for the usual circumstance in which more than one sitting must be arranged; the problem, of course, is compounded by the need for posttests. In September, 1981 an additional stimulus had to be rushed into use at Guelph because three sittings (consecutive evenings) were required, plus a later sitting for students who missed the original testing period.

An additional problem with writing in the expository - point of view/ argument mode is that topics inevitably become stale.

Here we recommend immediate development and screening of at least two further writing stimuli, and preferably more.

3. Earlier Testing

The question inevitably arises (and recent proposals by the University of Toronto underline the question): can tests such as these be used as admission tests? And the clear answer is that they could. A very important policy issue as well as a number of other issues emerges when this question is raised. We have noted earlier some of the problems that arise concerning the accuracy of measurement. And that is only one important problem among many in the area of entrance testing.

Once the subtests are made available by COU, COU itself will have to consider limitations and cautions.

Even if the battery is restricted to those users for which it was designed, there are advantages in providing, with the co-operation of universities,

schools systems, and perhaps the twinned Ministries, occasions of administration much earlier than registration week with the attendant problems noted in Section V. The Advisory Committee was, on the whole, positively disposed to the provision of earlier testing occasions in the secondary school setting.

A number of options are possible. Three of these are suggested (on the assumption that a number of universities may wish to make use of the results):

(1) universal administration in mid-year in the student's home school, (i.e. available for students wishing to take the test in view of specific post-secondary plans);

(2) administration in examination centres across the province on three occasions over the calendar year, with scores promptly returned to students so that students with unsatisfactory scores could undertake further appropriate English courses while still in secondary school and have an opportunity to take the test again; (the opportunity would exist for students in Grade 11 up, and mature students, to write the test when they believe they are ready, and the student could forward his/her score to the university with his/her application);

(3) a variation on either (1) or (2): the university could also administer the test in registration week for only those students (e.g. out of province) who so far have not had opportunity to take the test, and for those who to date had received unsatisfactory scores.

4. Availability of the OTEA Tests

Whereas we strongly recommend that COU make the test as widely available as possible for *bona fide* users in Ontario, we also strongly recommend that the protection of test confidentiality be a consideration when developing a procedure for the release of the tests.

5. Further Contractual Development Recommended

(1) Development of a Student Handbook to complement the battery.

(2) Development of two further alternate forms for Language, Thesis Development, and Reading and at least two further writing stimuli. (Plus calibration of all four forms of the tests.)

(3) Analysis (with report, sample compositions) of a sample of the writing in response to Forms A and B - September administration.

APPENDIX A

POSTTESTING WRITTEN ENGLISH

In the context where it is the intent to posttest only those students selected for a remedial or special composition course, unique problems emerge. These students would normally be chiefly low scorers in the initial (September) testing in which their writing was assessed comparatively with the writing of all the test takers.

Because in holistic scoring comparative judgment is inevitable in assigning a score or relative ranking, the posttesting of only the weaker writers will produce much higher scores and the illusion, therefore, of far more improvement than is the case. Thus, a piece of writing that in September would have received, say, a 3 on the initial 1-10 scale might receive a 6 or 7 in December or April simply because there is no representation from the better writers who are not being posttested.

Another problem is the topic variable. Students being posttested presumably are now writing on a different topic than they were in the initial administration.

The "shifting scale" problem is the more serious. In the 1981-82 posttesting we attempted to control for it by burial for scoring by all scorers of a number of September essays in the 125-130 range and a number which fell around 150, the mean score. Where these essays, re-scored, now fell would, we hoped, serve as rough cutting points to categorize writers who did/did not rise to a competence level in the posttest administration, and those who rose/did not rise to a level of competence above the mean September performance.

The experiment, unfortunately, was not a success because there were too few essays -- hence too few scorers -- for us to obtain sufficiently reliable new scores for these buried division-point essays.

We would anticipate that those universities interested in posttesting only "treatment" students identified initially as weaker writers will encounter the same problem.

To overcome the problem using holistic scoring where only such students are re-writing is very difficult. It would require burial in first-round administration of sufficient essays to guarantee that several of these would fall at roughly these points (*ca.* 15th percentile and 50th percentile) and their re-scoring in the posttest administration by a substantial number of scorers -- at least five.

We suggest the following procedures instead:

1. In posttesting the writing <u>where only the less able writers are being tested</u>, a small team be gathered to classify the writing as follows:

    1. Unacceptable
    2. Marginally acceptable (Here scores on the other subtests used in posttesting might compensate)
    3. Acceptable
    4. Superior

At least two scorers would independently make a rapid "holistic" judgment of the essay, with a third employed to referee discrepancies such as those between 1 and 3.

Prior to the scoring, a number of essays reasonably representative of the quality range should be selected for close analysis and discussion with reference to agreed-upon criteria such as basic organization, quality and sufficiency of argument, and mechanics. A reasonable consensus should be reached and the criteria "internalized" before the rapid scoring begins. The team might wish to make use of the cutting-point essays appended to the main report in establishing a division between categories 1 and 2.

2. Where it is possible to re-test the weaker group together with a fresh population undergoing initial testing, the problem does not emerge with such severity. The essays can be scored holistically with the others. This procedure is, wherever it is possible, preferable.

3. Another way to achieve an effect similar to that produced by including the essays of retested students undergoing initial testing would be to include the essays of retested students among a standard set of essays. In this case, the standard set would be selected from the essays written by an unselected group of students required to write the test. The distribution of scores for the standard set should span the range of the score scale.

APPENDIX B

REPORT ON SCORER RELIABILITY

As noted in the main report, the holistic scoring procedure includes the selection of ten essays per topic scored that are then inserted in every scorer's set. The scores from the ten essays are used to calibrate the scores for all the essays scored. The following tables give the reliability of individual scorers on the calibration essays for the pretest in three settings. The value for R is the correlation over essays between marks assigned by a particular scorer and the mean of all scores assigned to the essays.

Table 1: Laurentian University Pretest - Form B

Scorer No.	R	N	Essay No. 1	2	3	4	5	6	7	8	9	10
1	0.69	10	5	6	5	7	10	4	8	6	1	2
2	0.87	9	-	4	6	7	9	4	8	7	3	10
3	0.88	10	6	2	7	6	8	6	10	5	1	9
4	0.91	10	8	7	7	7	9	5	9	7	2	8
5	0.77	10	9	4	6	5	7	7	8	8	1	10
6	0.92	10	9	3	4	7	8	2	10	6	1	5
7	0.98	10	7	3	7	7	9	4	10	6	1	8
8	0.87	10	8	1	6	7	9	4	10	6	2	3
9	0.95	10	8	3	6	8	7	4	8	6	1	6

Essay #10, which "behaved" strangely in scorers' hands, was recommended for deletion in calibration.

The R for most scorers was remarkably high, and quite acceptable for all.

Table 2: University of Guelph Pretest - Forms A, B and C

A	B	R	(N)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
2.313	0.518	0.766	30	6	2	1	6	5	6	2	8	7	5	7	8	8	3	2	5	9	5	3	4	7	8	3	6	6	1	10	4	6	3
2.550	0.548	0.719	20	3	2	1	0	7	2	2	0	0	6	5	4	4	4	0	0	0	0	2	0	7	6	0	6	6	1	0	5	2	7
1.899	0.590	0.712	30	6	6	1	5	4	8	7	4	7	8	7	7	8	5	1	5	5	7	4	6	5	7	2	5	5	2	6	5	3	7
1.362	0.739	0.892	30	3	2	2	5	4	7	5	7	9	4	5	7	6	4	3	5	4	6	2	6	7	6	3	5	4	1	9	6	5	6
2.245	0.517	0.577	10	8	4	2	0	0	6	4	0	0	3	5	7	4	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.324	0.690	0.784	30	5	5	2	4	7	8	3	6	8	8	7	7	6	4	5	7	3	7	2	7	6	6	5	4	6	1	6	4	5	6
1.889	0.728	0.906	10	3	3	1	0	0	7	2	0	0	5	6	6	4	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3.327	0.419	0.676	10	0	0	0	5	0	0	0	6	10	0	0	0	0	0	2	5	1	8	0	2	0	0	5	0	0	0	6	0	0	0
0.771	0.717	0.798	20	0	0	0	5	4	0	0	8	7	0	0	0	0	0	3	6	5	7	7	6	7	7	5	7	7	1	8	7	6	9
2.496	0.636	0.905	10	0	0	0	2	0	0	0	5	8	0	0	0	0	0	1	7	2	5	0	6	0	0	3	0	0	0	7	0	0	0
2.087	0.667	0.705	10	0	0	0	2	0	0	0	4	6	0	0	0	0	0	3	7	4	5	0	6	0	0	5	0	0	0	8	0	0	0
1.739	0.657	0.694	10	0	0	0	4	0	0	0	8	8	0	0	0	0	0	4	5	7	4	0	6	0	0	3	0	0	0	7	0	0	0
2.300	0.574	0.852	20	0	0	0	3	4	0	0	6	9	0	0	0	0	0	2	7	6	8	2	6	6	4	1	6	4	1	10	5	5	4
2.274	0.471	0.860	10	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	6	10	0	7	5	1	0	9	3	8
1.639	0.769	0.718	10	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	5	3	0	4	5	1	0	6	3	6

Read zero as a blank. Some scorers scored one, some two, and some all three topics.

Again, conformity is quite reasonable given that three topics were involved and that actual scoring of the essays occurred from time to time over the following two weeks or so (unlike the Laurentian situation, where most were scored within a couple of days).

For only one scorer did R fall below 0.6; all but two met or exceeded 0.7.

It may be worth noting this performance especially as a closer approximation to most universities' situations where a sizable number of essays must be scored over time.

Table 3: Wilfrid Laurier University Pretest - Form B

WHO	A	B	R	(N)	1	2	3	4	5	6	7	8	9	10
1	2.938	0.477	0.929	10	10	4	5	3	2	8	7	6	1	9
2	2.646	0.443	0.696	10	9	6	8	4	2	7	8	6	5	10
3	5.349	0.038	0.075	10	4	8	2	10	9	3	6	7	1	5
4	1.822	0.779	0.776	10	6	5	4	4	2	6	5	6	3	7
5	3.694	0.316	0.555	10	5	3	4	5	9	8	9	7	1	8
6	3.589	0.358	0.698	10	10	3	4	8	5	2	7	6	1	9
7	3.118	0.509	0.813	10	6	3	7	4	1	5	8	6	1	7
8	2.655	0.434	0.396	10	8	4	6	5	8	7	8	8	6	7
9	2.802	0.484	0.881	10	10	3	7	5	4	9	6	4	1	8
10	4.170	0.257	0.284	10	8	4	4	6	8	5	7	4	4	4
11	3.164	0.444	0.470	10	6	5	6	5	3	6	4	5	2	7
12	2.679	0.497	0.914	10	10	5	4	3	4	7	8	7	1	9
13	3.746	0.293	0.545	10	10	9	8	8	3	9	4	4	2	5
14	3.204	0.491	0.939	10	7	2	4	3	3	5	9	6	1	3
15	3.989	0.303	0.489	10	6	9	3	4	5	9	5	4	1	5
16	1.762	0.673	0.828	10	7	5	5	5	4	8	8	7	2	5
17	-1.018	1.218	0.548	10	6	6	5	5	4	5	6	6	5	6
18	2.675	0.515	0.719	10	8	5	6	3	2	5	8	4	3	7
19	2.418	0.616	0.710	10	6	7	5	6	2	7	5	5	2	6
20	3.246	0.445	0.712	10	8	3	3	6	7	6	8	3	1	7
21	1.821	0.656	0.772	10	7	6	7	3	5	7	7	7	2	6
22	3.700	0.326	0.642	10	8	2	9	6	1	7	4	7	3	10
23	3.407	0.384	0.721	10	8	3	9	6	3	4	5	7	1	10
24	2.349	0.669	0.697	10	5	3	5	5	5	5	8	4	2	6
25	3.649	0.347	0.677	10	10	7	3	8	1	9	5	4	2	6
26	3.331	0.376	0.572	10	8	7	6	4	2	4	6	6	1	2
27	1.938	0.614	0.778	10	7	5	7	5	7	6	6	7	1	8
28	3.420	0.389	0.758	10	9	5	6	3	2	10	7	3	1	4
29	1.951	0.523	0.717	10	10	8	8	5	6	5	9	3	3	7
30	3.468	0.343	0.514	10	7	3	5	8	4	10	6	4	5	9

Scorer reliability here leaves much to be desired; five of the thirty fall below 0.5 for R, and five more fall between 0.5 and 0.6.

The evidence here suggests three cautions that may be recommended:

1. The training session is very important.
2. Some reasonable time must be allowed for scoring.
3. Multiple scoring is essential in fairness to students and, as well, it is desirable to monitor scoring. With a large number of scorers perhaps in different locations it is probably infeasible to conduct a checking/refereeing/rescoring system during the process, but a *post facto* correction is possible through deletion of scores from scorers such as #3, #8 and #10 in the preceding exhibit. Perhaps, needless to add, the run identifies scorers who should not be re-employed.

APPENDIX C

DESCRIPTIVE STATISTICS FOR THE SUBTESTS

1. Form A (N=1698)	Mean (%)	S.D.	Reliability
II Reading	51.78	3.77	0.60
III Thesis Development	61.10	3.12	0.58
IV Language	48.93	5.41	0.71
2. Form B (N=357)			
II Reading	51.56	3.59	0.63
III Thesis Development	61.45	3.17	0.60
IV Language	45.60	5.37	0.72

3. Reliability Estimates for Combined Subtests II-IV:

Form A	0.80
Form B	0.83

4. I: Written English

As stated in the main report, in every setting the essay scores were scaled 100-200 with a mean of 150 and a standard deviation of approximately 20.